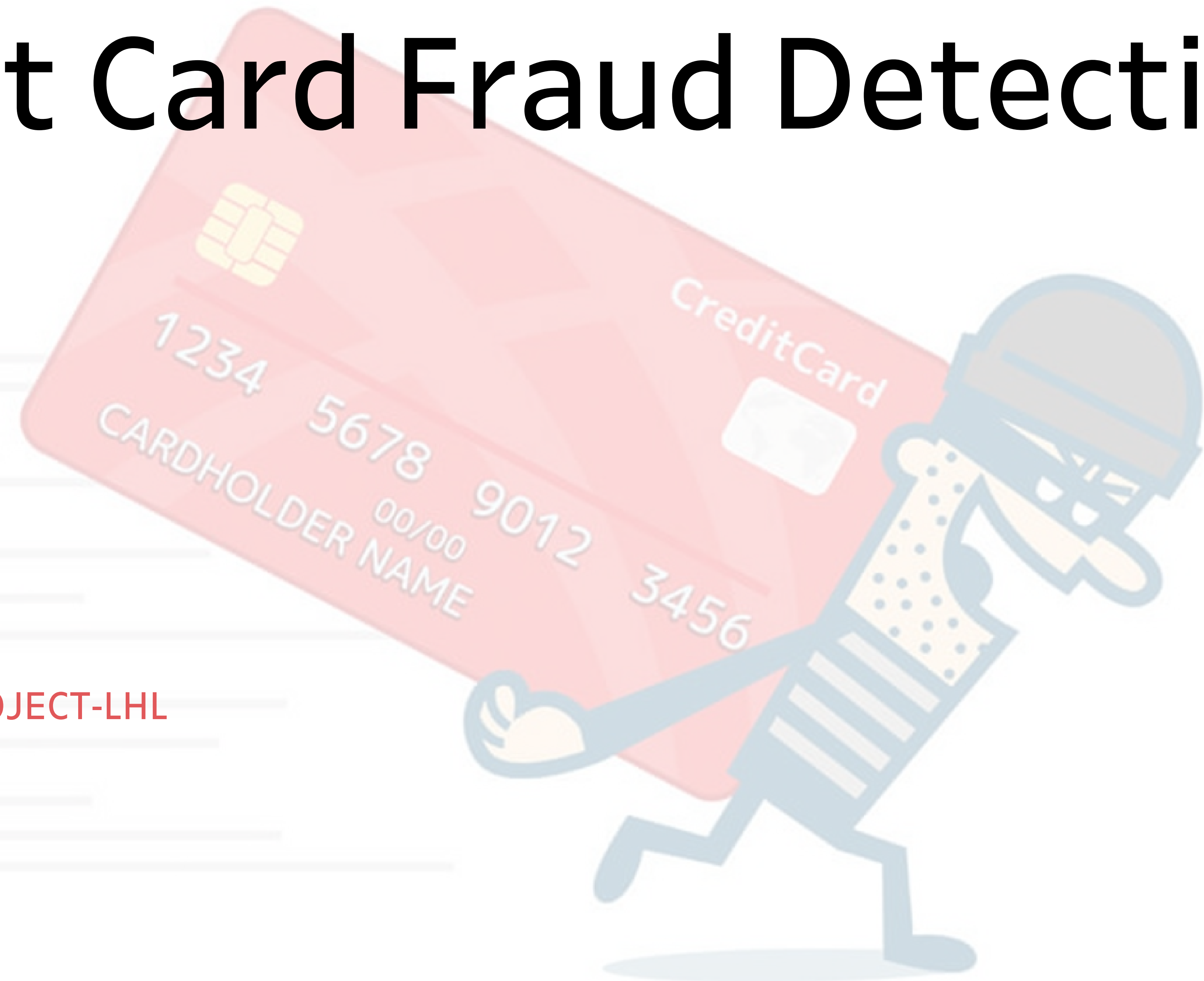
Credit Card Fraud Detection



DS MIDTERM PROJECT-LHL

By

Shaief Wares
Ahmad Furqan

Overview:

- The main focus of the project is to provide insight into credit card fraud by conducting exploratory data analysis and building visualizations and dashboards to make observations and trends, such as:
    - Distribution of transaction amounts between fraudulent and legitimate transactions
    - Transactions based on time of day
    - Geographical patterns of transactions
    - Correlation of certain variables and occurence of fraud
    - Distribution of spending amounts for fraudulent and legitimate transactions
    - Differences in distribution of variables between fraudulent and legitimate transactions
    - Whether or not certain merchants are more susceptible to fraud than others

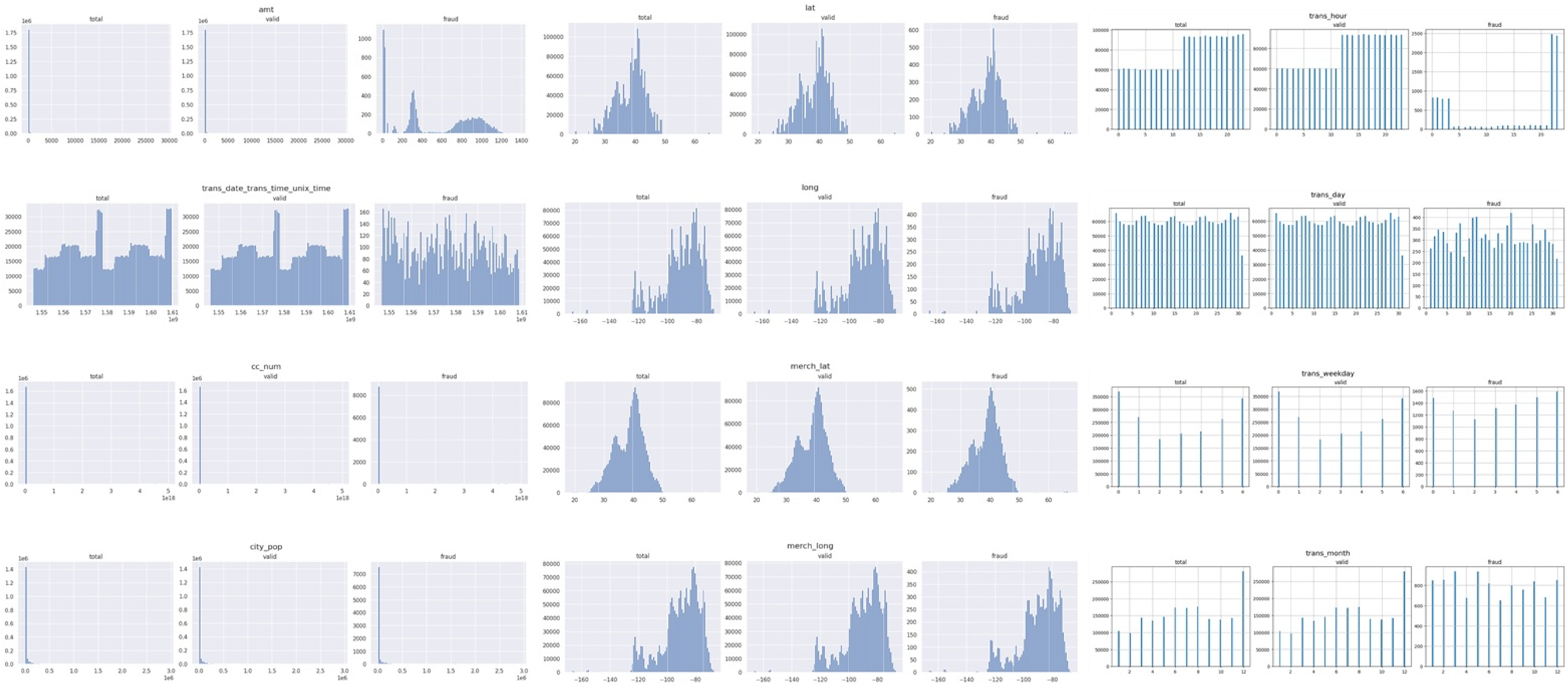
- Secondary objective is to look into predicting credit card fraud by utilizing the data set for model building and forecasting (including evaluating the built model)

The Dataset/EDA:

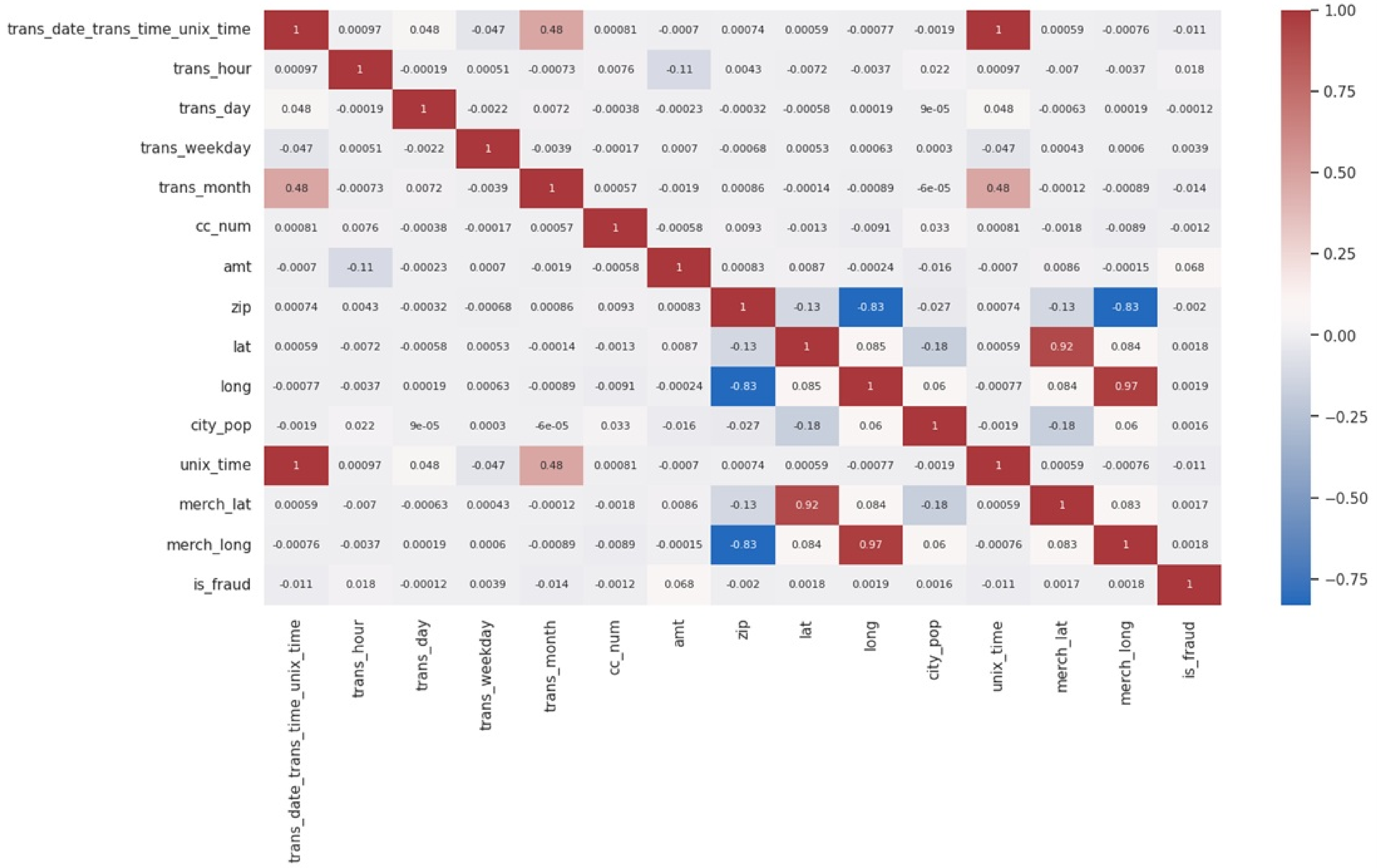
- This dataset is comprehensive log of simulated credit card transactions with over 1.85 million entries, providing a realistic and diverse environment (and thus valuable) for our investigation.
- The transactions span from January 1st, 2019 to December 31st, 2020 and encompasses both legitimate and fraudulent transactions.
-The dataset features transactions by 1,000 customers in over 900 cities all across the US engaging with around 700 distinct merchants.
- The purchases are categorized in 14 different categories and the amount spent ranges from as little as $1 to as much as $29,000.
- Only 9651 of the 1.85 million transactions (0.52%) were fraudulent.

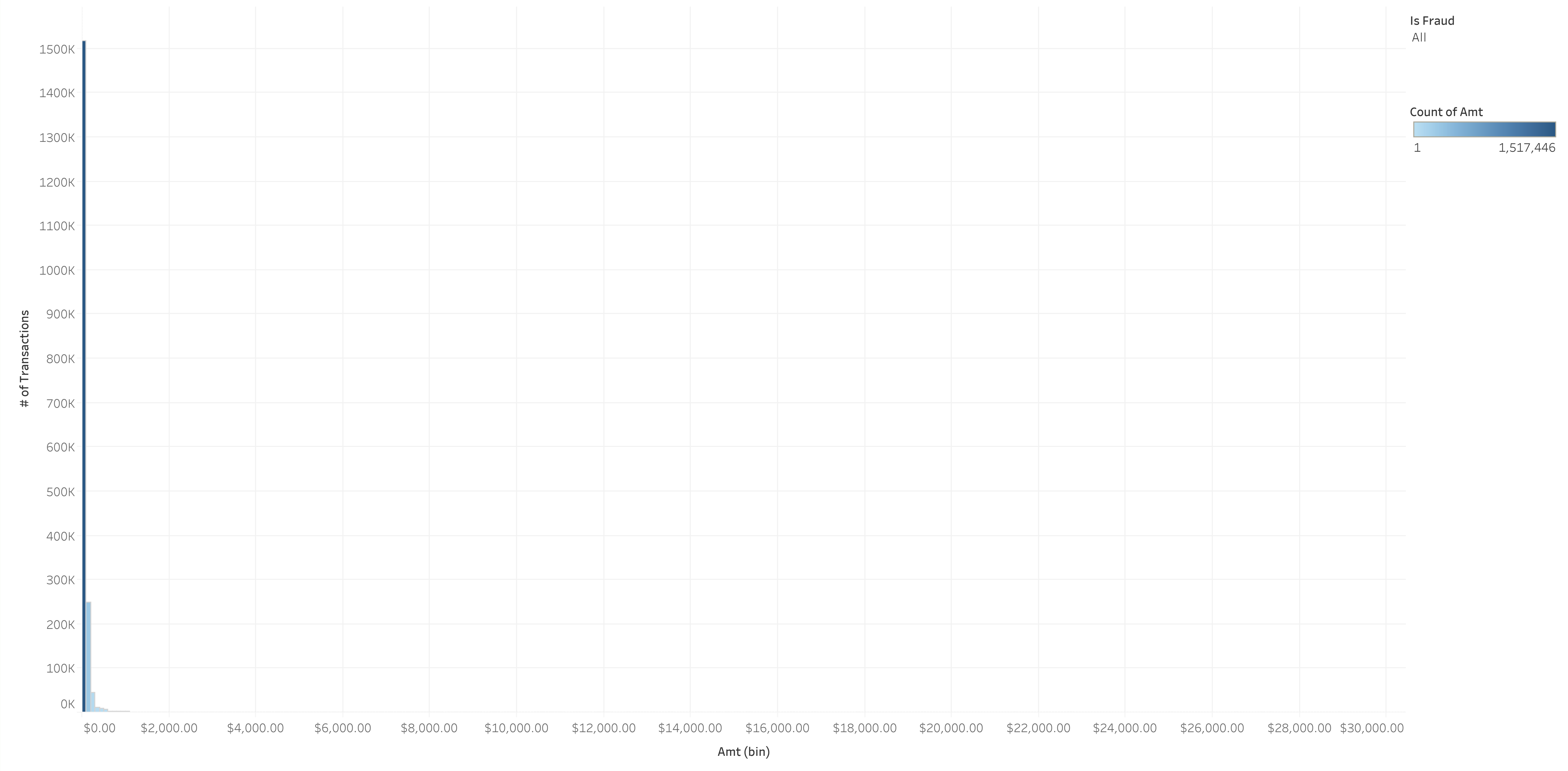
- The dataset includes the following features:
  - **Timestamps:** Each transaction is timestamped, enabling time-based pattern analysis
  - **Merchant Details:** Name of merchant and geolocations (i.e. latitude & longitude)
  - **Transaction Categories:** Category of the type of product purchased
  - **Transaction Amount:** How much was spent for each transaction
  - **Credit Card Holder Information:** Names, gender, dates of birth, addresses, CC numbers
  - **Fraud Indicator (is_fraud):** A binary flag (1 for fraudulent transactions, 0 for legitimate)

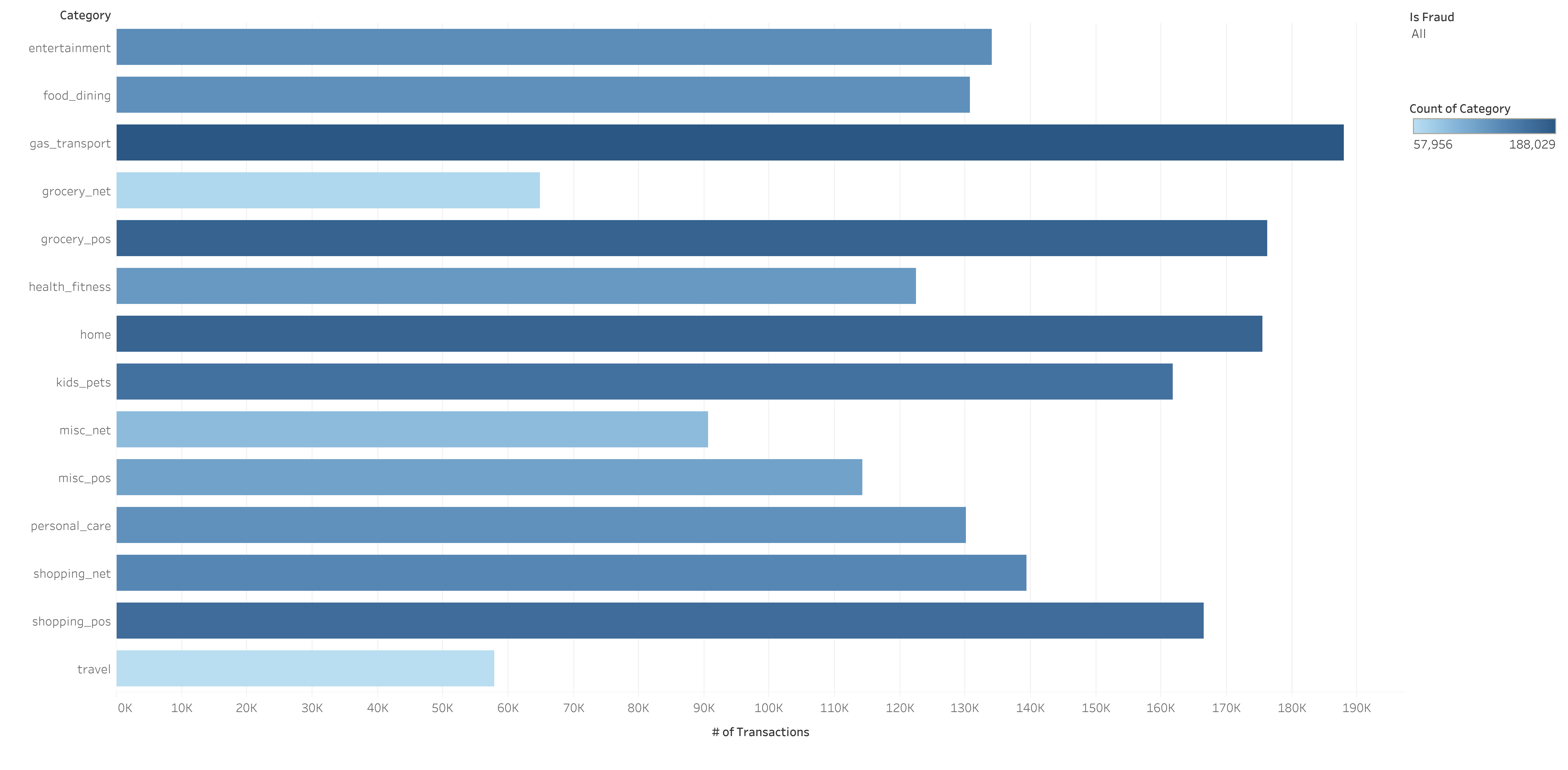
EDA (Histograms)



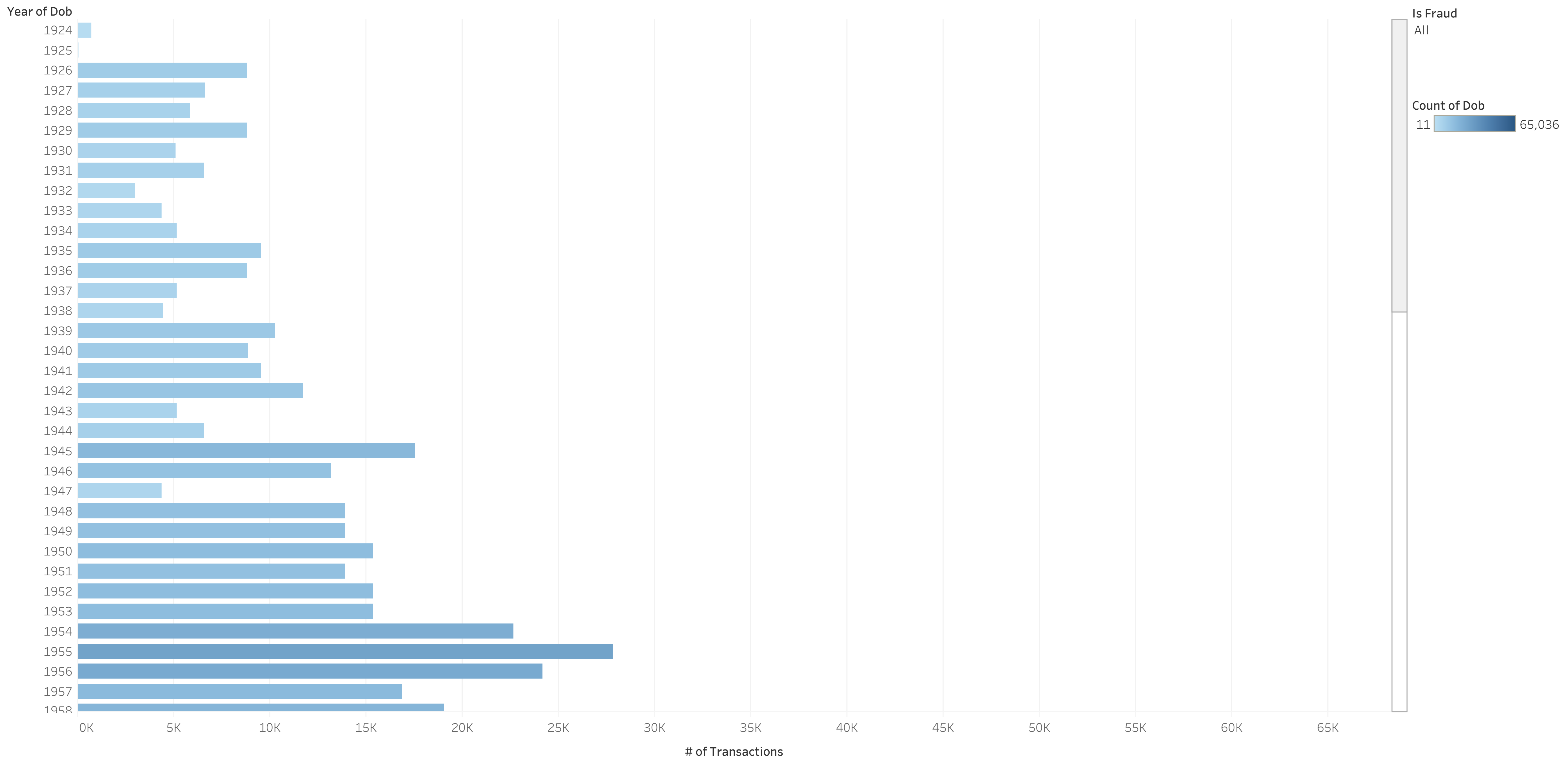
EDA (Correlation Heatmap)



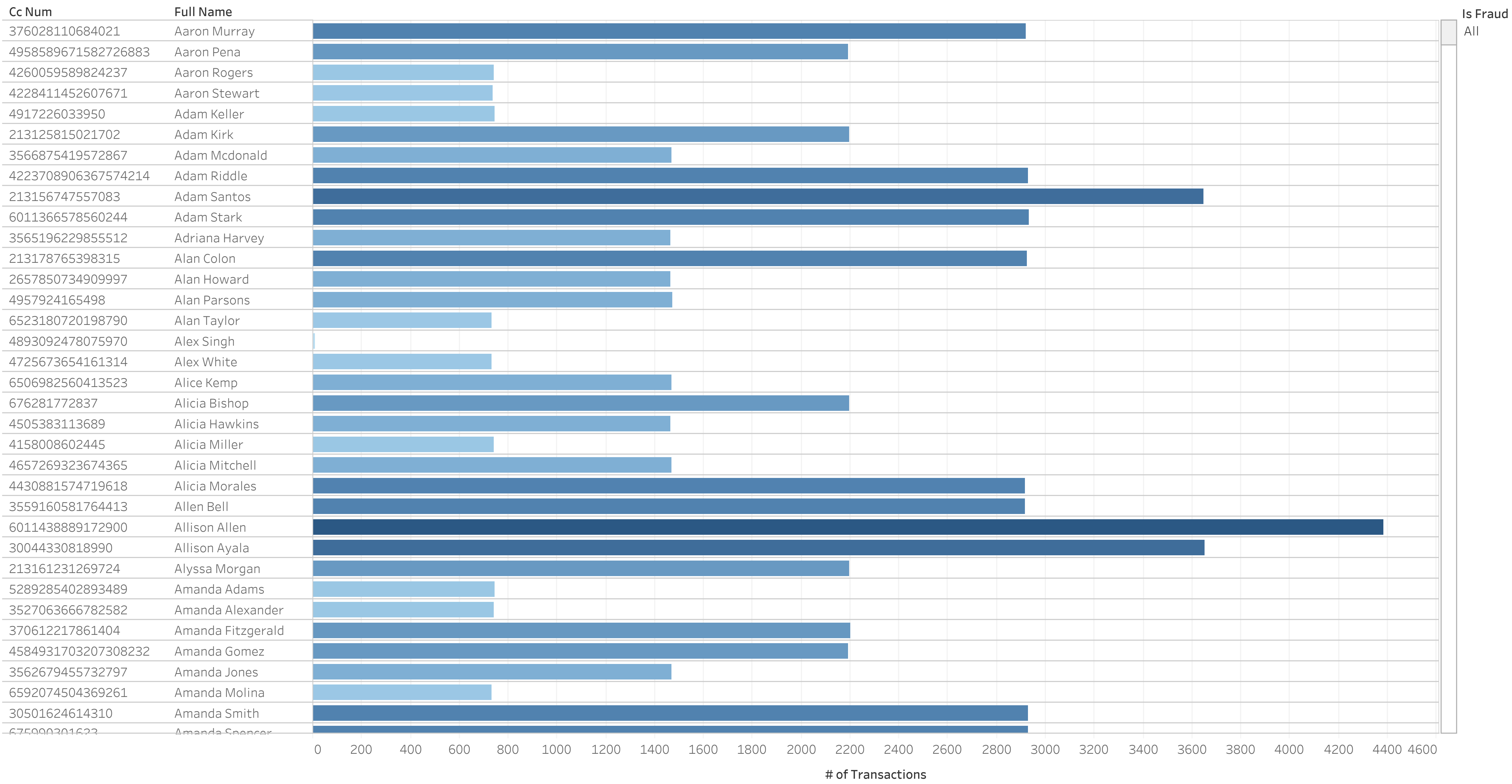

Is Fraud
All
Count of Amt
1
1,517,446
# of Transactions
1500K
1400K
1300K
1200K
1100K
1000K
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
$0.00
$2,000.00
$4,000.00
$6,000.00
$8,000.00
$10,000.00
$12,000.00
$14,000.00
$16,000.00
$18,000.00
$20,000.00
$22,000.00
$24,000.00
$26,000.00
$28,000.00
$30,000.00
Amt (bin)


Category
entertainment
food_dining
gas_transport
grocery_net
grocery_pos
health_fitness
home
kids_pets
misc_net
misc_pos
personal_care
shopping_net
shopping_pos
travel
0K
10K
20K
30K
40K
50K
60K
70K
80K
90K
100K
110K
120K
130K
140K
150K
160K
170K
180K
190K
# of Transactions
Is Fraud
All
Count of Category
57,956
188,029

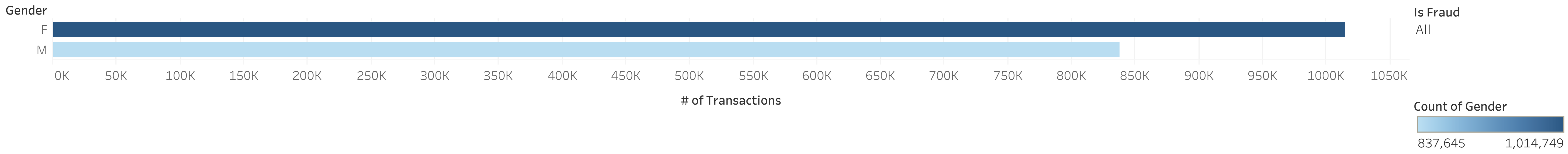
Dataset | Dataset (Continued) | EDA (Correlation Heatmap) | Fraud By Amount | Fraud Categories | Fraud by Year of Birth | Fraud by Credit Card Numbers | Fraud by Gender | Fraud by Occupation | Fraud By State | Fraud By Merchant



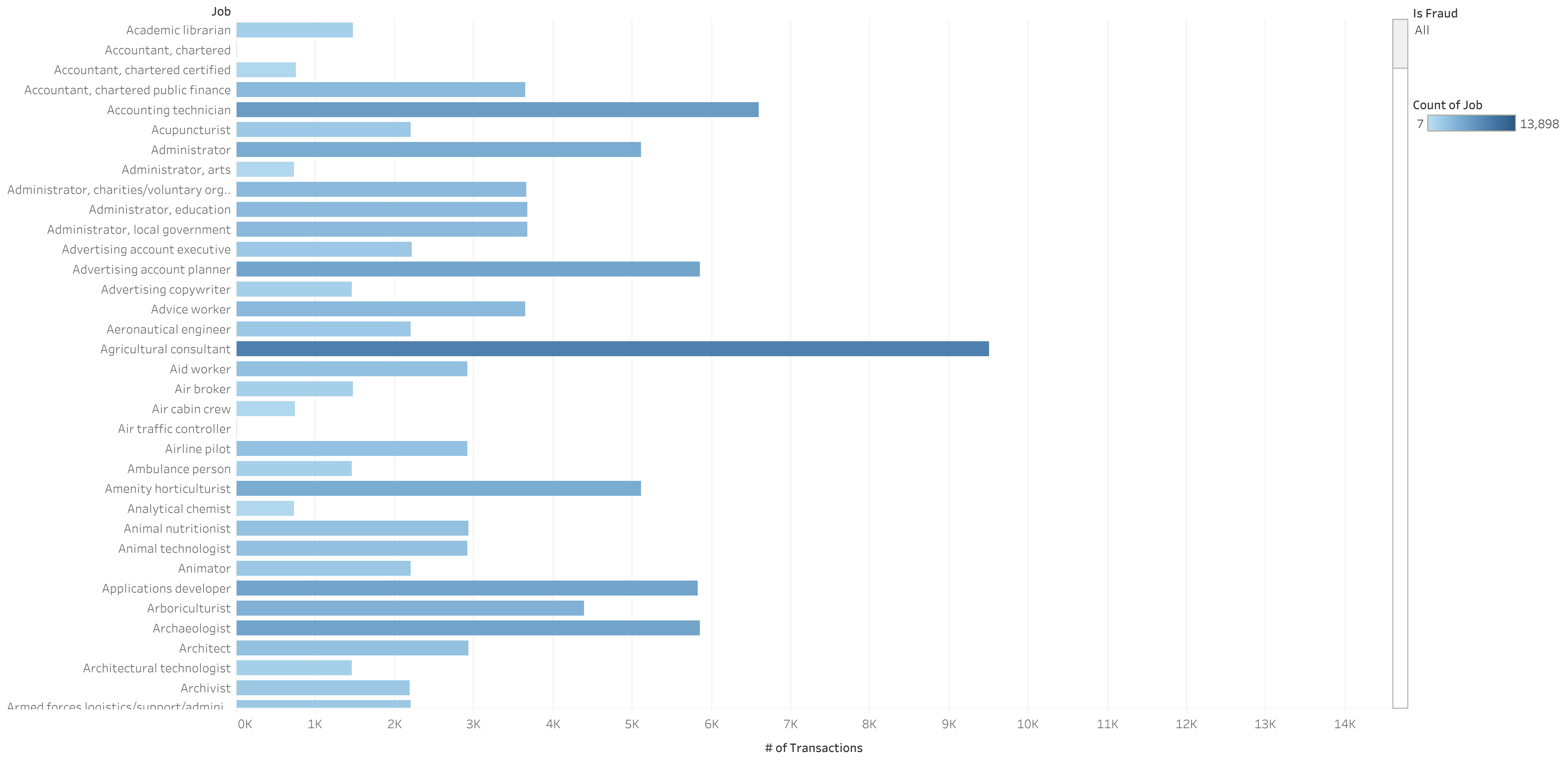
Dataset (Continued) | EDA (Correlation Heatmap) | Fraud By Amount | Fraud Categories | Fraud by Year of Birth | Fraud by Credit Card Numbers | Fraud by Gender | Fraud by Occupation | Fraud By State | Fraud By Merchant | Individual Dashboard



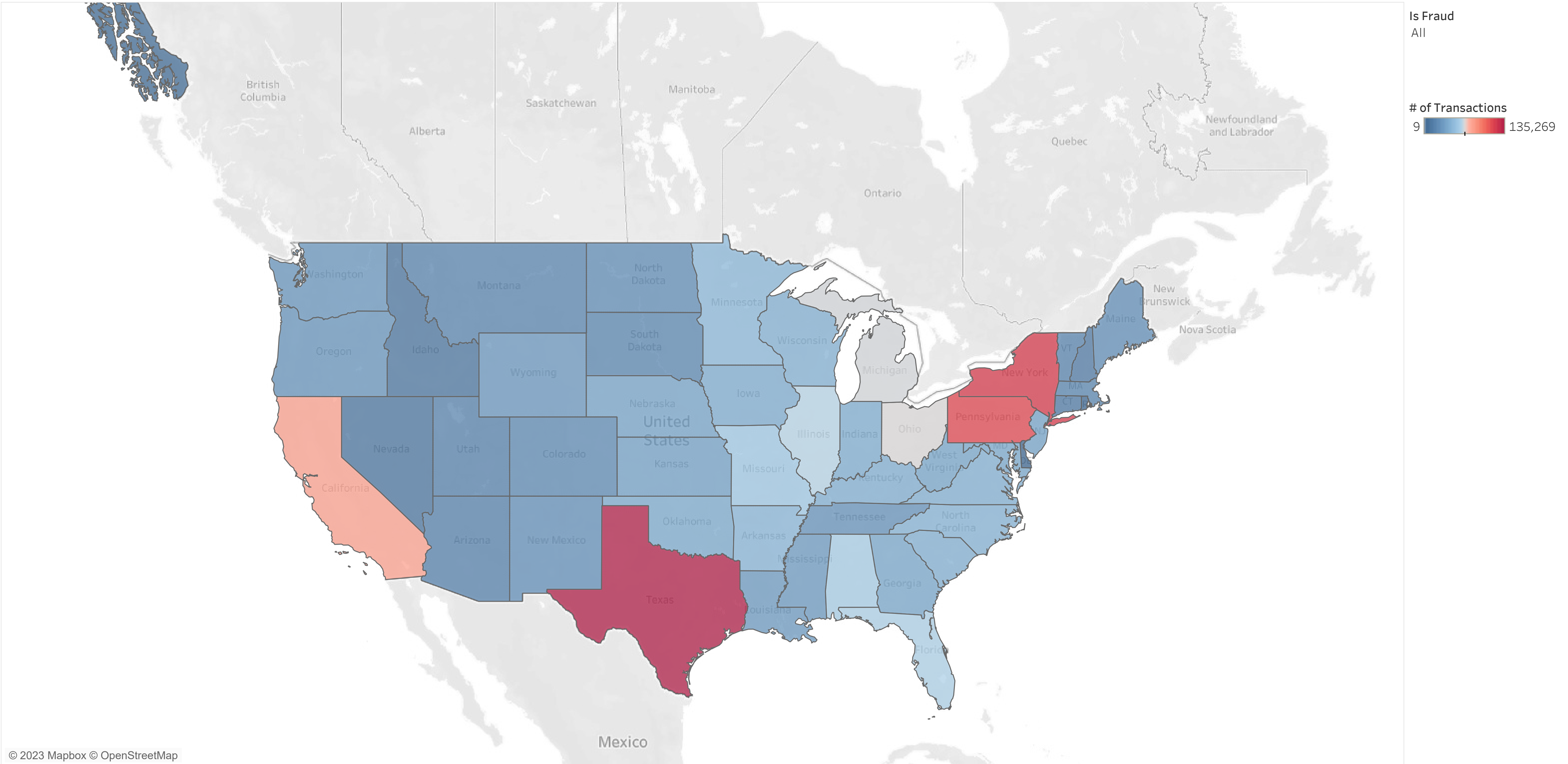
EDA (Correlation Heatmap)	Fraud By Amount	Fraud Categories	Fraud by Year of Birth	Fraud by Credit Card Numbers	Fraud by Gender	Fraud by Occupation	Fraud By State	Fraud By Merchant	Individual Dashboard	Occupation, Year of Birth, and Gender Dashboard



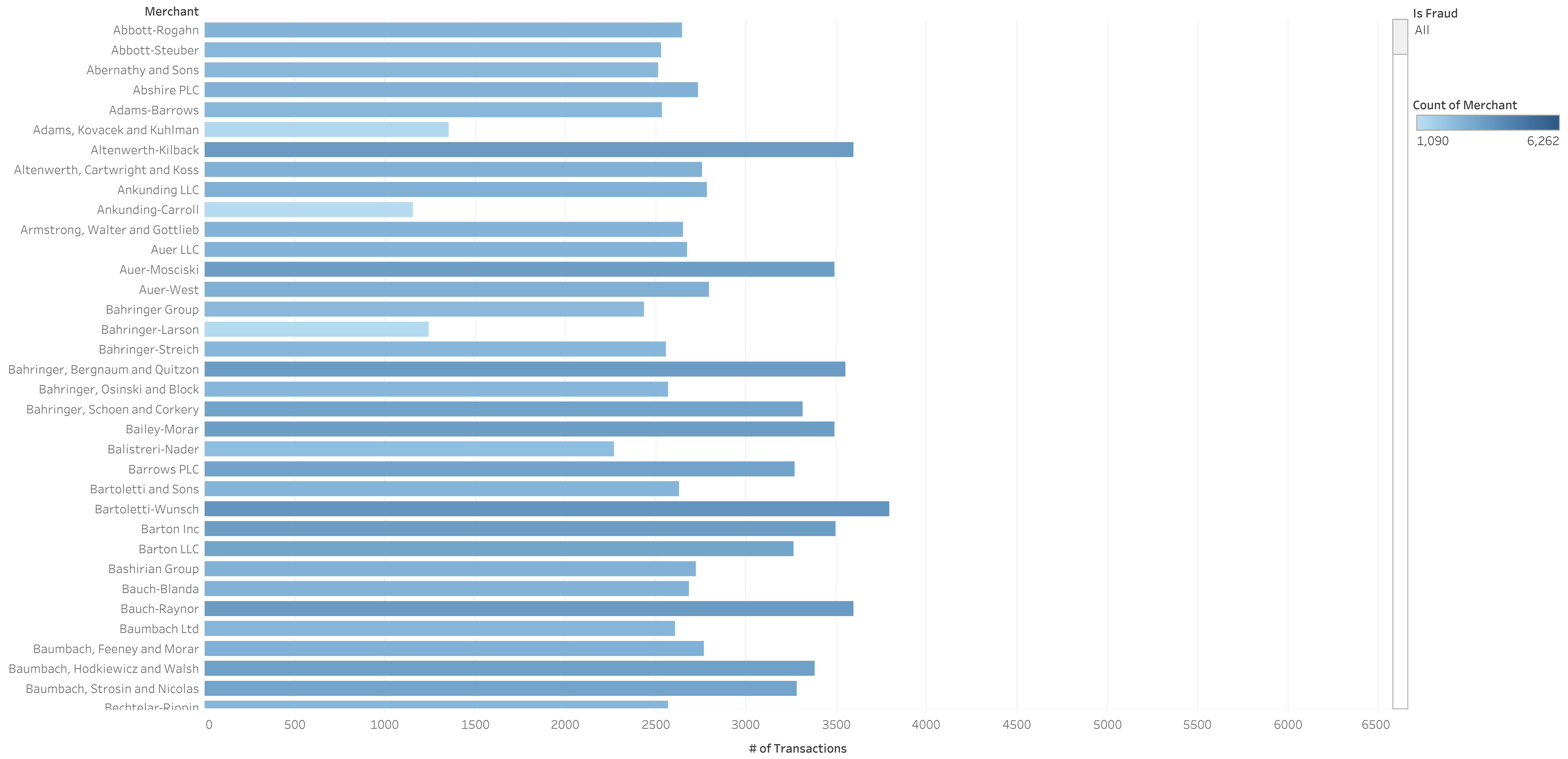
Fraud By Amount | Fraud Categories | Fraud by Year of Birth | Fraud by Credit Card Numbers | Fraud by Gender | Fraud by Occupation | Fraud By State | Fraud By Merchant | Individual Dashboard | Occupation, Year of Birth, and Gender Dashboard | Merchant Dashboard



Fraud Categories	Fraud by Year of Birth	Fraud by Credit Card Numbers	Fraud by Gender	Fraud by Occupation	Fraud By State	Fraud By Merchant	Individual Dashboard	Occupation, Year of Birth, and Gender Dashboard	Merchant Dashboard	Credit Card Dashboard



Fraud by Year of Birth | Fraud by Credit Card Numbers | Fraud by Gender | Fraud by Occupation | Fraud By State | **Fraud By Merchant** | Individual Dashboard | Occupation, Year of Birth, and Gender Dashboard | Merchant Dashboard | Credit Card Dashboard | Location Dashboard



Fraud by Credit Card Numbers | Fraud by Gender | Fraud by Occupation | Fraud By State | Fraud By Merchant | Individual Dashboard | Occupation, Year of Birth, and Gender Dashboard | Merchant Dashboard | Credit Card Dashboard | Location Dashboard | Time Dashboard

Aaron Murrav

Aaron Murray

Gender: M

Date of Birth: December 23, 1974

Occupation: Tourist information centre manager

Address:
624 Hale Springs Apt. 572
Meadville, MO 64659

Credit Card(s):
376028110684021

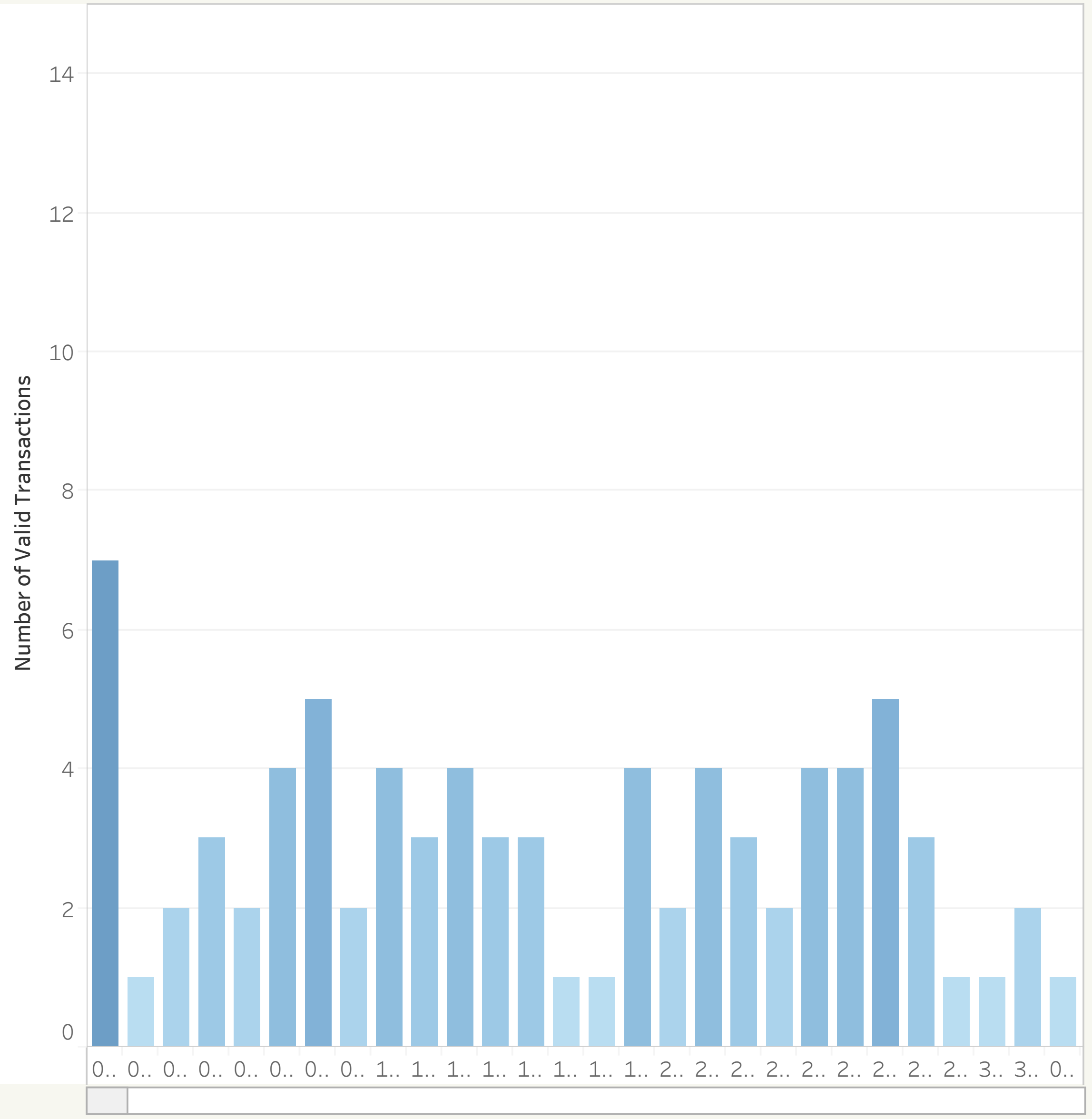
Total Number of Transactions: 2,920

Valid Transactions: 2,912 **Fraudulent Transactions:** 8

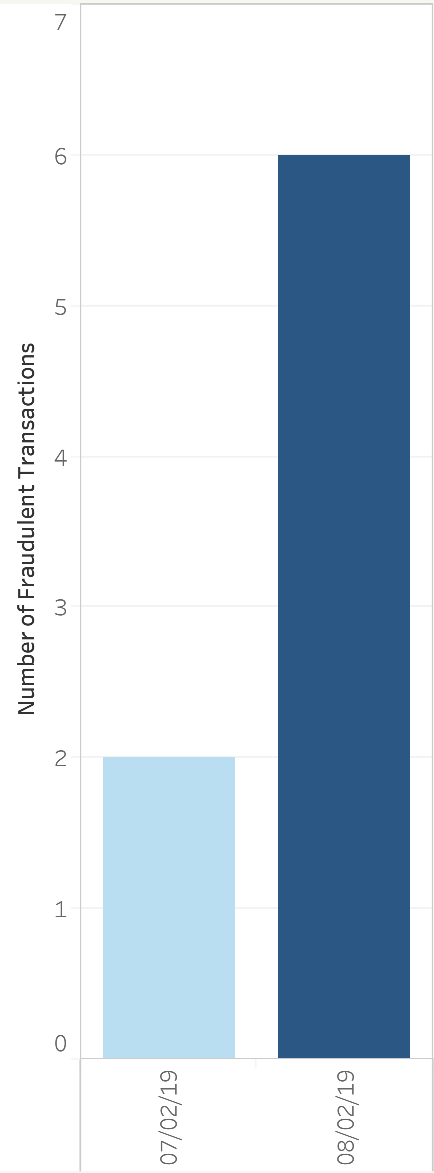
Total Spent: $286,923

Total Spent (Valid): $284,001 **Total Spent (Fraud):** $2,922

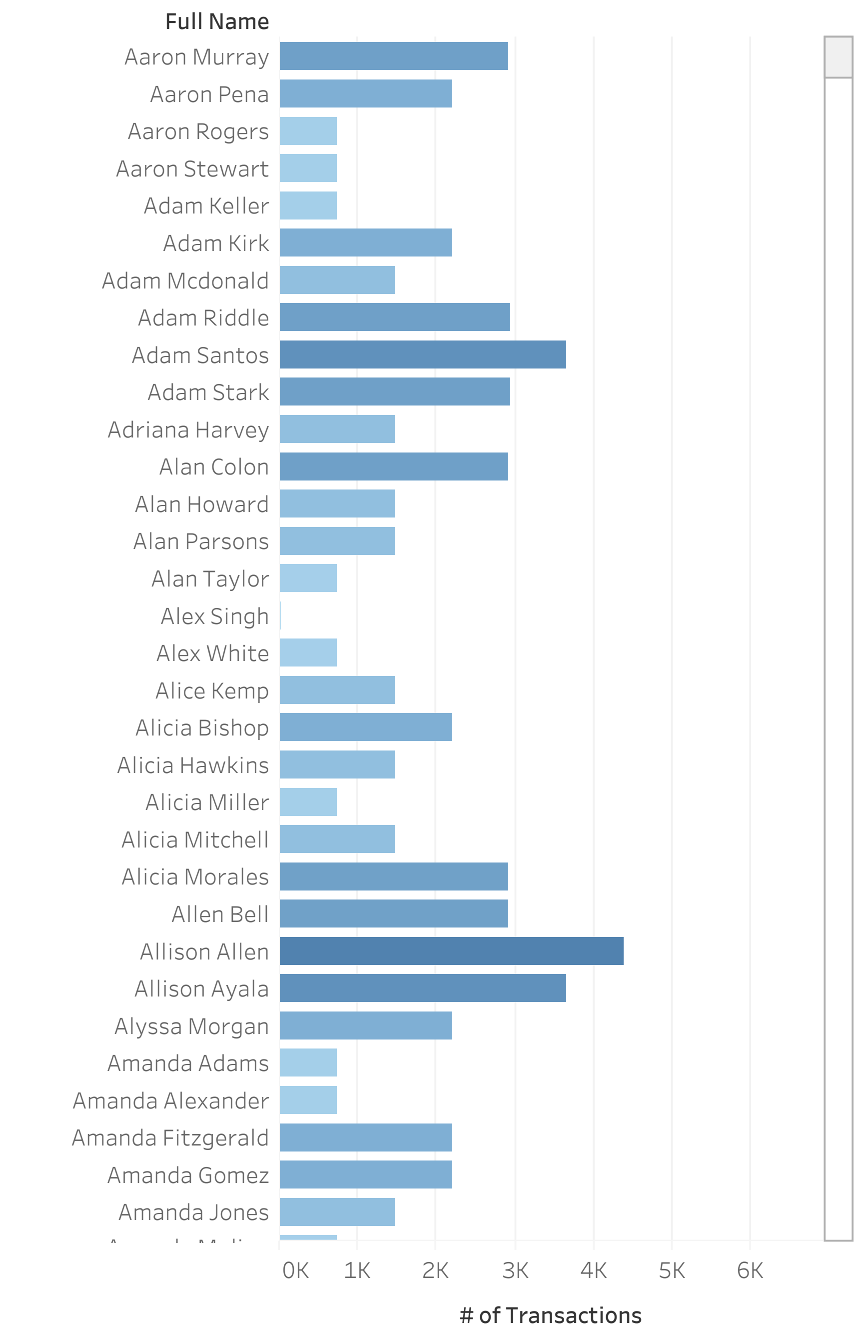
Transactions by Date (Valid)

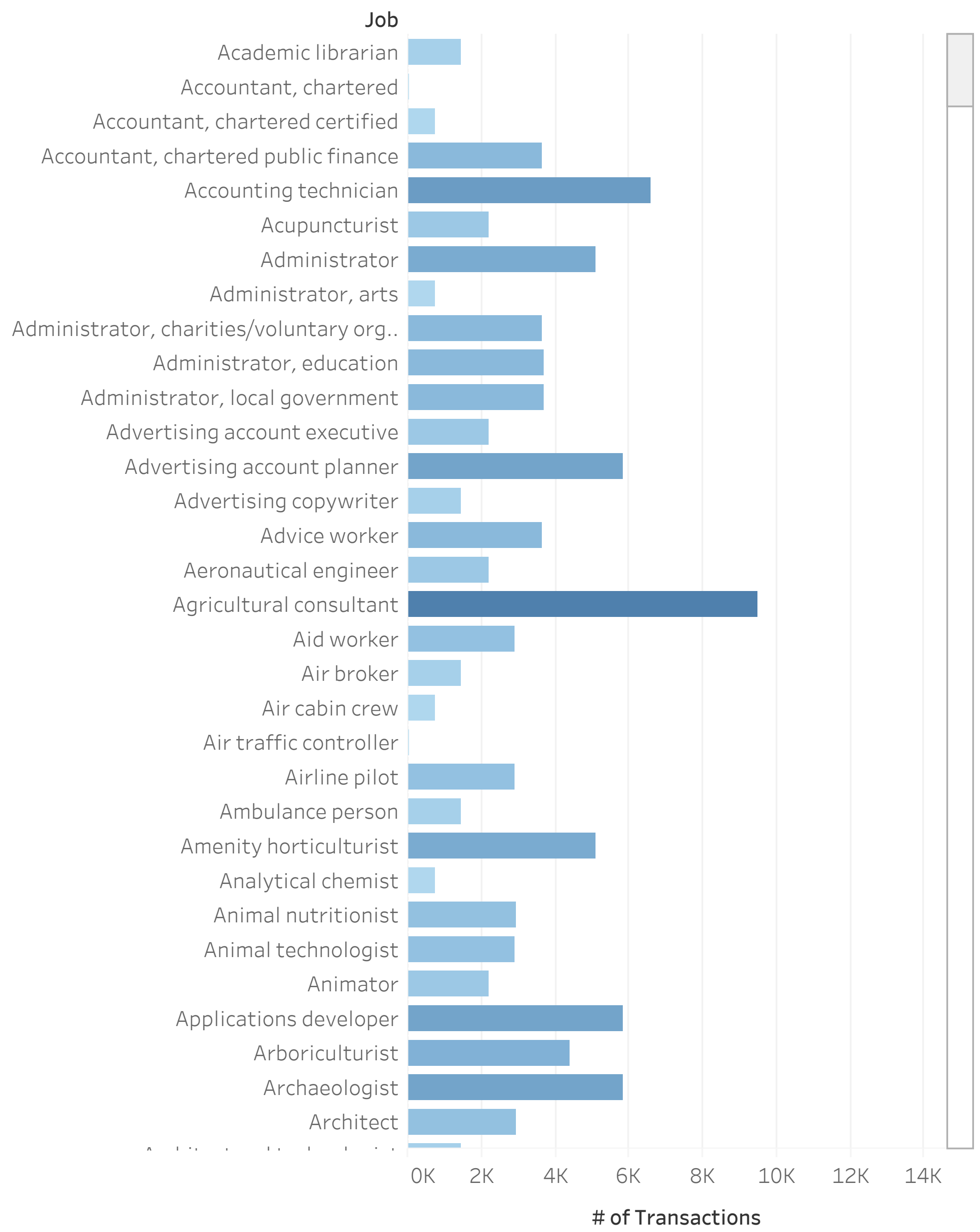


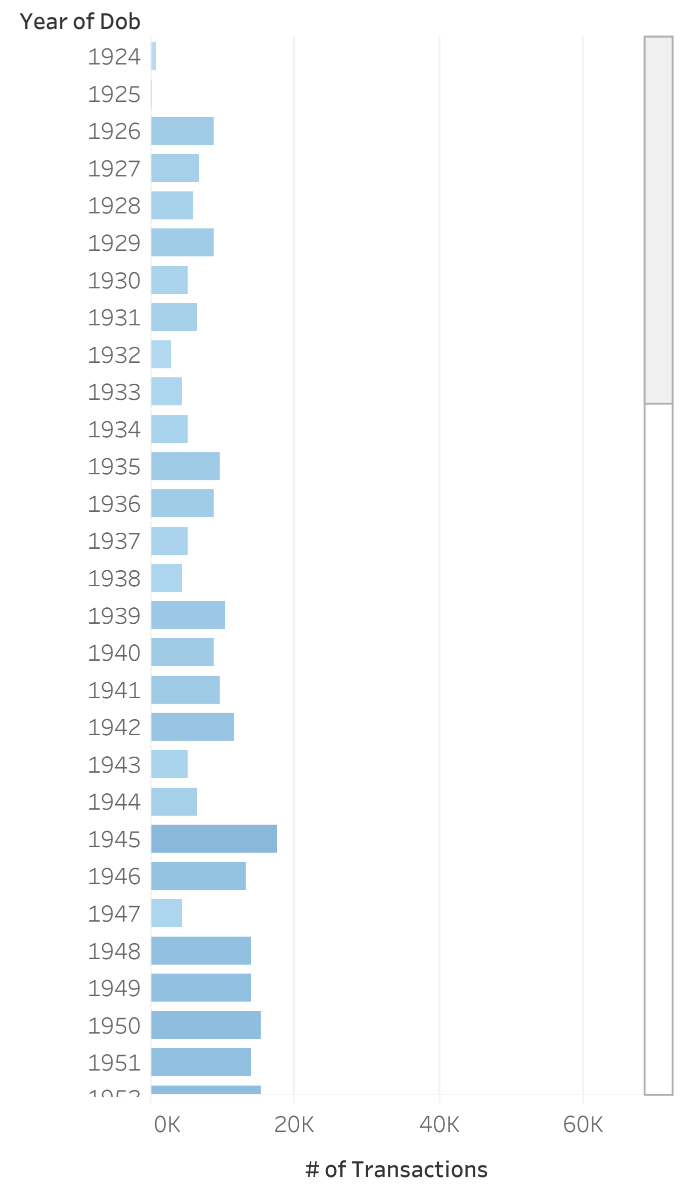
Transactions by Date (Fraud)



Fraud by Gender | Fraud by Occupation | Fraud By State | Fraud By Merchant | Individual Dashboard | Occupation, Year of Birth, and Gender Dashboard | Merchant Dashboard | Credit Card Dashboard | Location Dashboard | Time Dashboard | Modeling



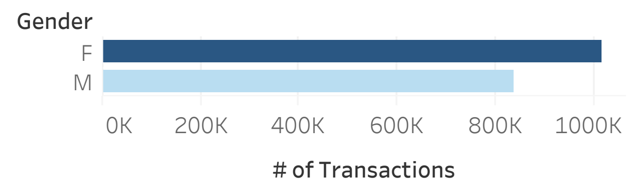




Occupation
All

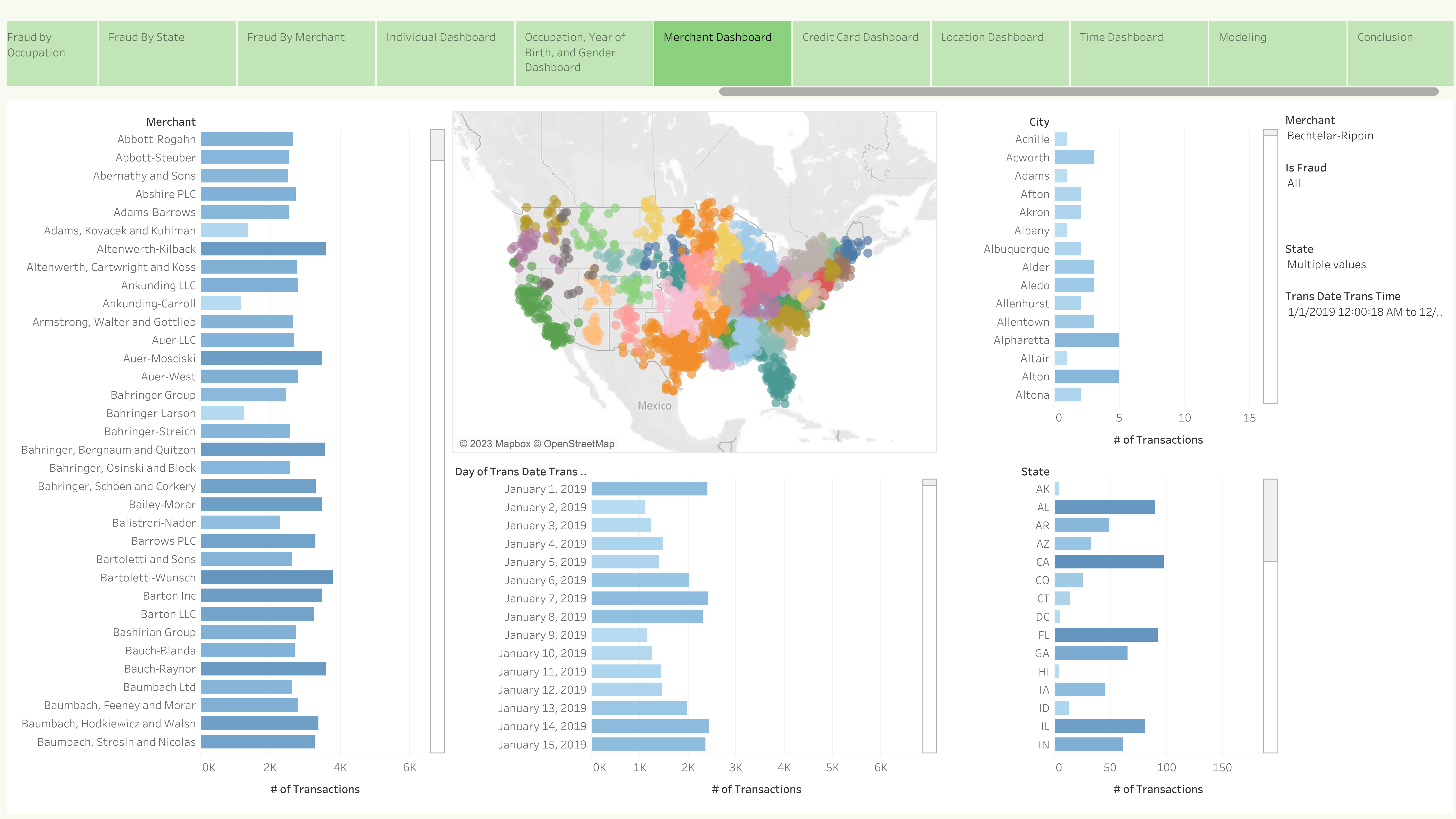
Birth Year
All

Is Fraud
All



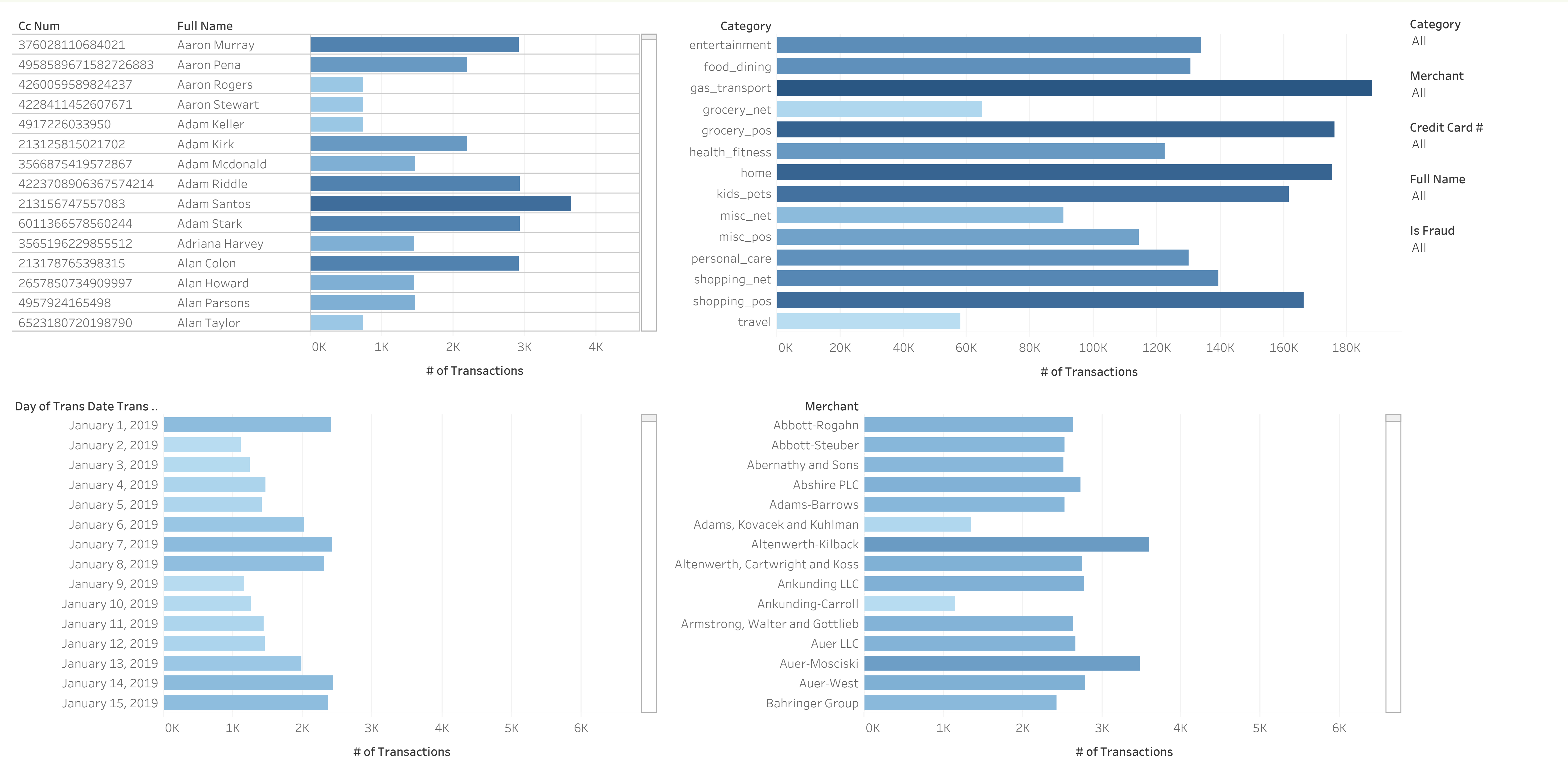
Total Number of Unique Individuals: 989

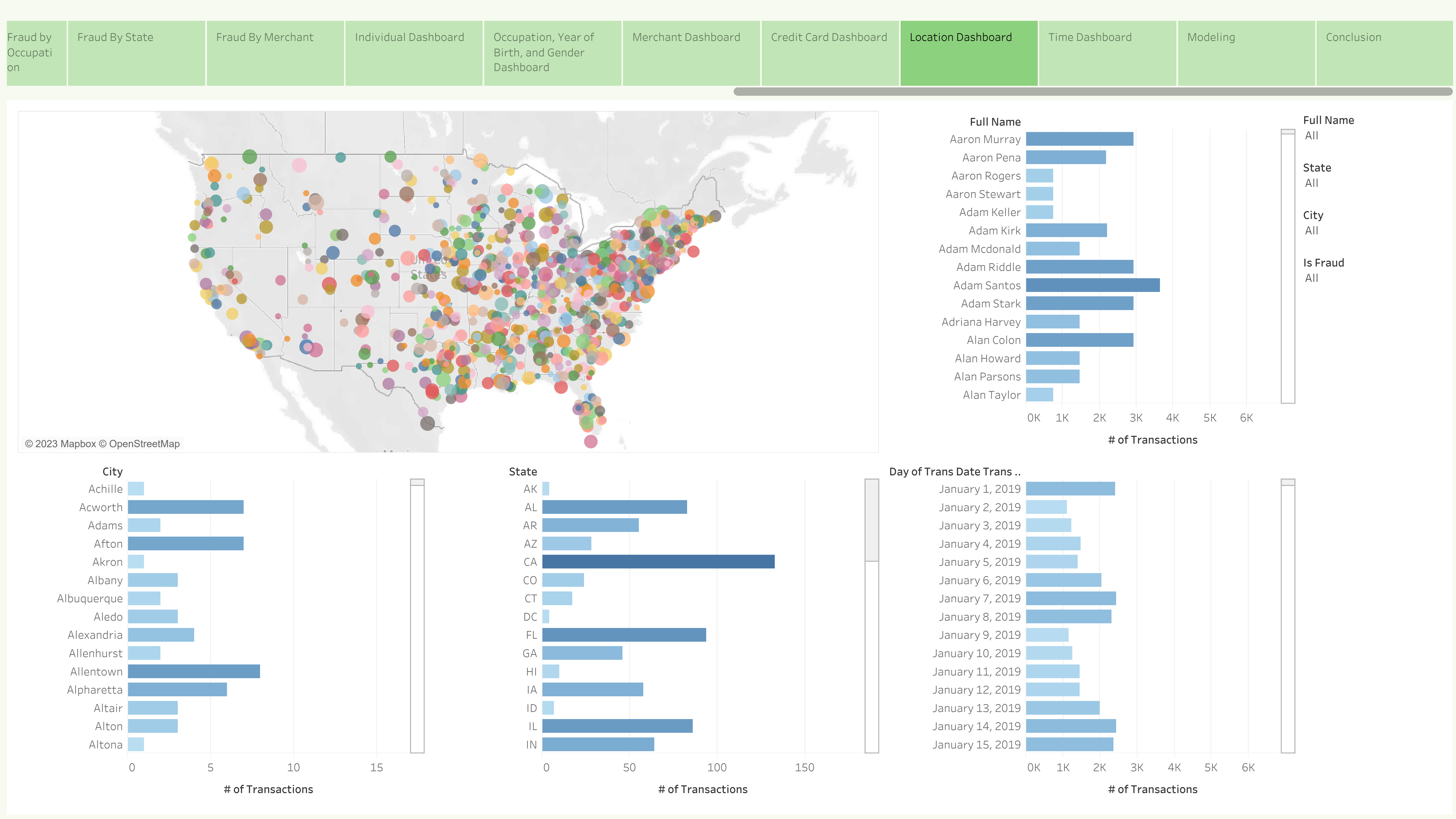
Total Number of Unique Occupations: 497



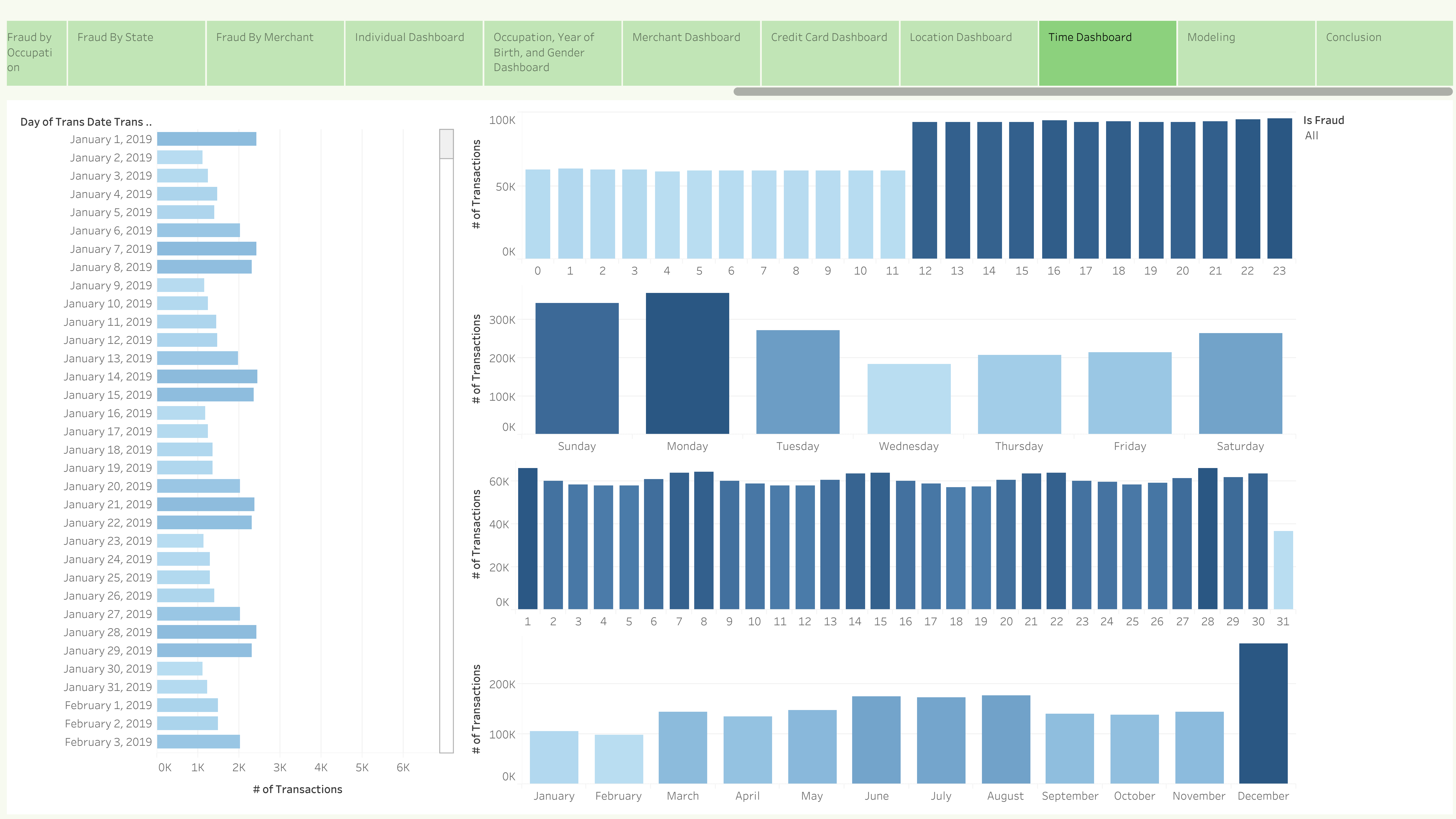
Fraud by Occupation
Fraud By State
Fraud By Merchant
Individual Dashboard
Occupation, Year of Birth, and Gender Dashboard
Merchant Dashboard
Credit Card Dashboard
Location Dashboard
Time Dashboard
Modeling
Conclusion
Merchant
Abbott-Rogahn
Abbott-Steuber
Abernathy and Sons
Abshire PLC
Adams-Barrows
Adams, Kovacek and Kuhlman
Altenwerth-Kilback
Altenwerth, Cartwright and Koss
Ankunding LLC
Ankunding-Carroll
Armstrong, Walter and Gottlieb
Auer LLC
Auer-Mosciski
Auer-West
Bahringer Group
Bahringer-Larson
Bahringer-Streich
Bahringer, Bergnaum and Quitzon
Bahringer, Osinski and Block
Bahringer, Schoen and Corkery
Bailey-Morar
Balistreri-Nader
Barrows PLC
Bartoletti and Sons
Bartoletti-Wunsch
Barton Inc
Barton LLC
Bashirian Group
Bauch-Blanda
Bauch-Raynor
Baumbach Ltd
Baumbach, Feeney and Morar
Baumbach, Hodkiewicz and Walsh
Baumbach, Strosin and Nicolas
0K
2K
4K
6K
# of Transactions
Mexico
© 2023 Mapbox © OpenStreetMap
Day of Trans Date Trans ..
January 1, 2019
January 2, 2019
January 3, 2019
January 4, 2019
January 5, 2019
January 6, 2019
January 7, 2019
January 8, 2019
January 9, 2019
January 10, 2019
January 11, 2019
January 12, 2019
January 13, 2019
January 14, 2019
January 15, 2019
0K
1K
2K
3K
4K
5K
6K
# of Transactions
City
Achille
Acworth
Adams
Afton
Akron
Albany
Albuquerque
Alder
Aledo
Allenhurst
Allentown
Alpharetta
Altair
Alton
Altona
0
5
10
15
# of Transactions
State
AK
AL
AR
AZ
CA
CO
CT
DC
FL
GA
HI
IA
ID
IL
IN
0
50
100
150
# of Transactions
Merchant
Bechtelar-Rippin
Is Fraud
All
State
Multiple values
Trans Date Trans Time
1/1/2019 12:00:18 AM to 12/..

Fraud by Occupation | Fraud By State | Fraud By Merchant | Individual Dashboard | Occupation, Year of Birth, and Gender Dashboard | Merchant Dashboard | Credit Card Dashboard | Location Dashboard | Time Dashboard | Modeling | Conclusion



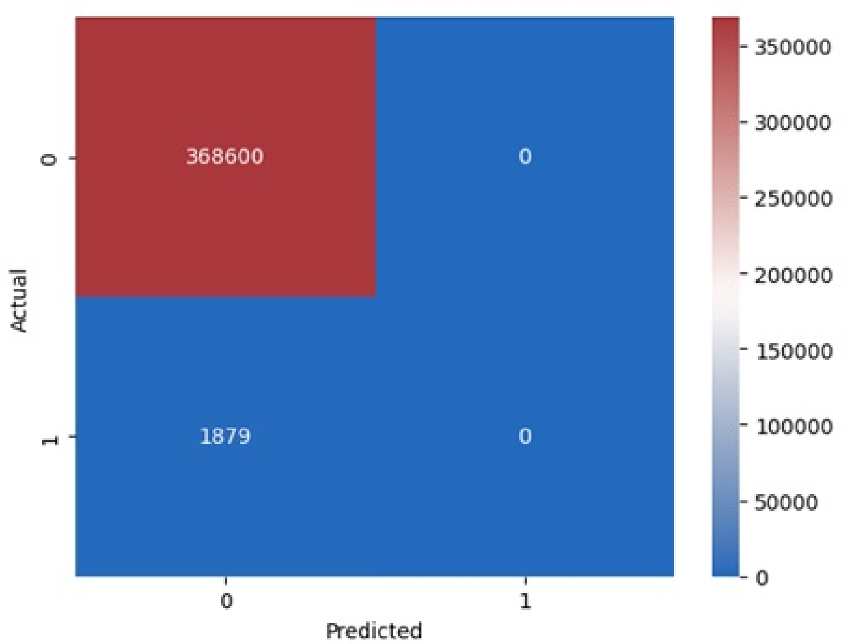


Fraud by Occupation
Fraud By State
Fraud By Merchant
Individual Dashboard
Occupation, Year of Birth, and Gender Dashboard
Merchant Dashboard
Credit Card Dashboard
Location Dashboard
Time Dashboard
Modeling
Conclusion
© 2023 Mapbox © OpenStreetMap
Full Name
Aaron Murray
Aaron Pena
Aaron Rogers
Aaron Stewart
Adam Keller
Adam Kirk
Adam Mcdonald
Adam Riddle
Adam Santos
Adam Stark
Adriana Harvey
Alan Colon
Alan Howard
Alan Parsons
Alan Taylor
0K 1K 2K 3K 4K 5K 6K
# of Transactions
Full Name
All
State
All
City
All
Is Fraud
All
City
Achille
Acworth
Adams
Afton
Akron
Albany
Albuquerque
Aledo
Alexandria
Allenhurst
Allentown
Alpharetta
Altair
Alton
Altona
0 5 10 15
# of Transactions
State
AK
AL
AR
AZ
CA
CO
CT
DC
FL
GA
HI
IA
ID
IL
IN
0 50 100 150
# of Transactions
Day of Trans Date Trans ..
January 1, 2019
January 2, 2019
January 3, 2019
January 4, 2019
January 5, 2019
January 6, 2019
January 7, 2019
January 8, 2019
January 9, 2019
January 10, 2019
January 11, 2019
January 12, 2019
January 13, 2019
January 14, 2019
January 15, 2019
0K 1K 2K 3K 4K 5K 6K
# of Transactions


Fraud by Occupation
Fraud By State
Fraud By Merchant
Individual Dashboard
Occupation, Year of Birth, and Gender Dashboard
Merchant Dashboard
Credit Card Dashboard
Location Dashboard
Time Dashboard
Modeling
Conclusion
Day of Trans Date Trans ..
January 1, 2019
January 2, 2019
January 3, 2019
January 4, 2019
January 5, 2019
January 6, 2019
January 7, 2019
January 8, 2019
January 9, 2019
January 10, 2019
January 11, 2019
January 12, 2019
January 13, 2019
January 14, 2019
January 15, 2019
January 16, 2019
January 17, 2019
January 18, 2019
January 19, 2019
January 20, 2019
January 21, 2019
January 22, 2019
January 23, 2019
January 24, 2019
January 25, 2019
January 26, 2019
January 27, 2019
January 28, 2019
January 29, 2019
January 30, 2019
January 31, 2019
February 1, 2019
February 2, 2019
February 3, 2019
0K 1K 2K 3K 4K 5K 6K
# of Transactions
# of Transactions
100K 50K 0K
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Is Fraud
All
# of Transactions
300K 200K 100K 0K
Sunday Monday Tuesday Wednesday Thursday Friday Saturday
# of Transactions
60K 40K 20K 0K
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
# of Transactions
200K 100K 0K
January February March April May June July August September October November December

Modeling:

- Initial use of logistic regression with scikit-learn.
- Evaluation using the confusion matrix showed no successful prediction of fraud cases.
- A second attempt was made with the logit function from the statsmodel library, which also employed logistic regression.
- The results summary indicated poor model performance, characterized by a notably low R-squared value.
- Despite efforts to improve the model by removing variables with high P-values, only marginal enhancement was achieved.




```
                           Logit Regression Results
==============================================================================
Dep. Variable:               is_fraud   No. Observations:            1481915
Model:                          Logit   Df Residuals:                1481902
Method:                           MLE   Df Model:                         12
Date:                Sun, 01 Oct 2023   Pseudo R-squ.:               0.09498
Time:                        23:25:27   Log-Likelihood:              -43947.
converged:                       True   LL-Null:                     -48559.
Covariance Type:            nonrobust   LLR p-value:                   0.000
==================================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           2.8644      1.183      2.422      0.015       0.547       5.182
trans_date_trans_time_unix_time -5.415e-09  7.54e-10     -7.184      0.000   -6.89e-09   -3.94e-09
trans_hour                      0.0215      0.002     12.141      0.000       0.018       0.025
trans_day                       0.0001      0.001      0.099      0.921      -0.002       0.003
trans_weekday                   0.0286      0.005      5.272      0.000       0.018       0.039
trans_month                    -0.0593      0.004    -14.876      0.000      -0.067      -0.052
cc_num                        -1.148e-20  9.28e-21     -1.237      0.216   -2.97e-20    6.71e-21
amt                             0.0024   2.85e-05     85.289      0.000       0.002       0.002
lat                             0.0396      0.021      1.917      0.055      -0.001       0.080
long                            0.0199      0.021      0.968      0.333      -0.020       0.060
city_pop                     5.307e-09   3.94e-08      0.135      0.893   -7.19e-08    8.25e-08
merch_lat                      -0.0319      0.021     -1.555      0.120      -0.072       0.008
merch_long                     -0.0181      0.021     -0.883      0.377      -0.058       0.022
==================================================================================================
                           Logit Regression Results
==============================================================================
Dep. Variable:               is_fraud   No. Observations:            1481915
Model:                          Logit   Df Residuals:                1481909
Method:                           MLE   Df Model:                          5
Date:                Sun, 01 Oct 2023   Pseudo R-squ.:               0.09478
Time:                        23:31:00   Log-Likelihood:              -43957.
converged:                       True   LL-Null:                     -48559.
Covariance Type:            nonrobust   LLR p-value:                   0.000
==================================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           2.9998      1.176      2.551      0.011       0.695       5.305
trans_date_trans_time_unix_time -5.414e-09  7.53e-10     -7.189      0.000   -6.89e-09   -3.94e-09
trans_hour                      0.0215      0.002     12.116      0.000       0.018       0.025
trans_weekday                   0.0286      0.005      5.272      0.000       0.018       0.039
trans_month                    -0.0593      0.004    -14.881      0.000      -0.067      -0.052
amt                             0.0024   2.85e-05     85.245      0.000       0.002       0.002
==================================================================================================
```

Conclusion:

Fraudulent transactions displayed a noticeable trend, primarily occurring during specific periods, typically a few hours before and after midnight.
These fraudulent transactions were characterized by notably lower spending amounts.
Many other transaction patterns resembled those of legitimate transactions.
However, due to the dataset's scarcity of fraudulent instances (i.e. an imbalanced dataset) and the intricate nature of the patterns, employing advanced modeling techniques like random forest or XGBoost is imperative for effective classification.